Machine Learning for Signal Processing
Sparse and Overcomplete Representations

Bhiksha Raj
(slides from Sourish Chaudhuri and Abelino Jimenez)

1

So far

Can we use linear composition to identify **basic units** that compose the signal?

2

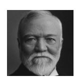
So far

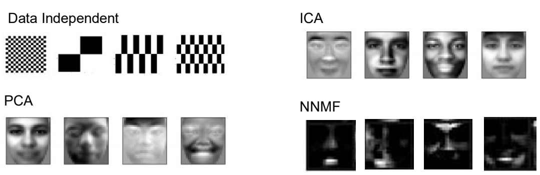
Weights

$$D \cdot \alpha = X \quad \approx$$

Basis

Data





basis <= dim X

3

Just in case you missed it..

- Remember, #(Basis Vectors)= #unknowns

$$D \cdot \alpha = X$$

Basis Vectors

Weights

Input data

Standard representations: number of bases <= dimension of data

4

A limitation we saw earlier

- Mathematical restrictions on the number of bases have no connection to reality
  - Universe does not respect your mathematical representations of the data
  - In reality: number of building blocks that compose any kind of data is unlimited
- One solution we saw earlier: picking *one* "closest" building block to represent any input

5

Poll 1

- Mark all true statements about the vector quantization model
  - It represents data as v = Dw where D is a dictionary
  - The w vector in vector-quantization is required to be one hot
  - K-means clustering is one way of computing the Dictionary for the VQ model
  - The Dictionary is assumed to represent semantically meaningful "bases" that can be used to compose the data
  - The Dictionary may be viewed as a collection of exemplars that all data instances are mapped onto
- What is the difference or similarity between VQ (Kmeans) based representations and KLT/PCA?
  - They are completely different concepts and cannot be compared
  - They are similar in that both of them minimize the L2 divergence between v and Dw
  - They differ in that in VQ w must be one-hot and the bases (dictionary) are unrestricted, whereas in KLT/PCA the bases (dictionary) must be orthogonal while w is unrestricted

6

Poll 1

- Mark all true statements about the vector quantization model
  - It represents data as v = Dw where D is a dictionary
  - The w vector in vector-quantization is required to be one hot
  - K-means clustering is one way of computing the Dictionary for the VQ model
  - The Dictionary is assumed to represent semantically meaningful "bases" that can be used to compose the data
  - The Dictionary may be viewed as a collection of exemplars that all data instances are mapped onto
- What is the difference or similarity between VQ (Kmeans) based representations and KLT/PCA?
  - They are completely different concepts and cannot be compared
  - They are similar in that both of them minimize the L2 divergence between v and Dw
  - They differ in that in VQ w must be one-hot and the bases (dictionary) are unrestricted, whereas in KLT/PCA the bases (dictionary) must be orthogonal while w is unrestricted

7

A limitation we saw earlier

- Mathematical restrictions on the number of bases have no connection to reality
  - Universe does not respect your mathematical representations of the data
  - In reality: number of building blocks that compose any kind of data is unlimited
- One solution we saw earlier: picking *one* "closest" building block to represent any input
- Today: Learning linear compositional representations without restrictions on the number of basic units

8

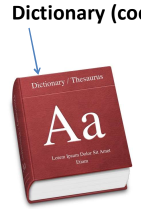
Key Topics in this Lecture

- Basics – Component-based representations
  - ***Overcomplete*** and Sparse Representations,
  - ***Dictionaries***
- Pursuit Algorithms
- How to learn a dictionary
- Why is an overcomplete representation powerful?

9

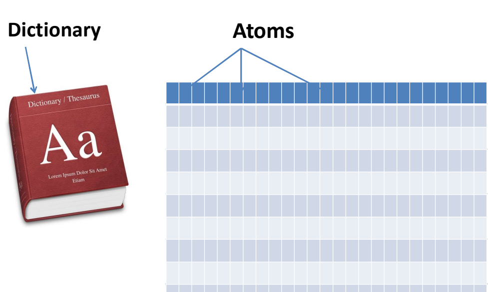
Representing Data

Dictionary (codebook)



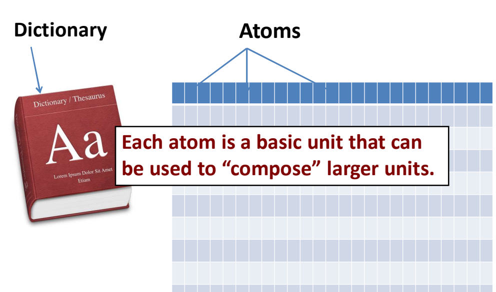
10

Representing Data

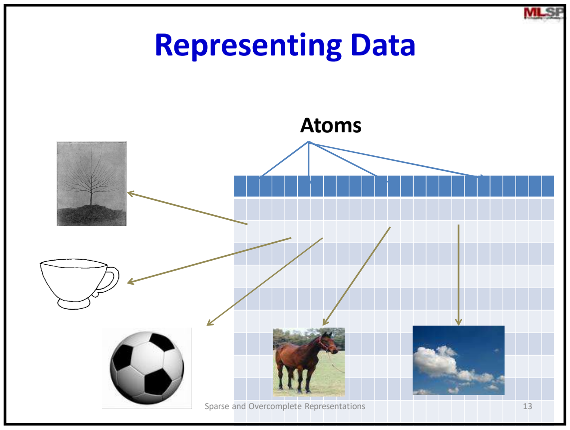


11

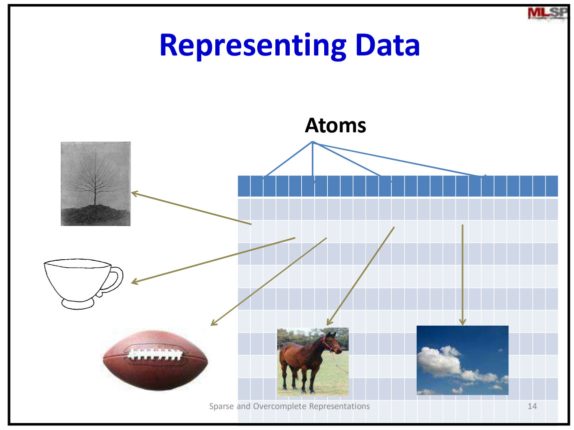
Representing Data



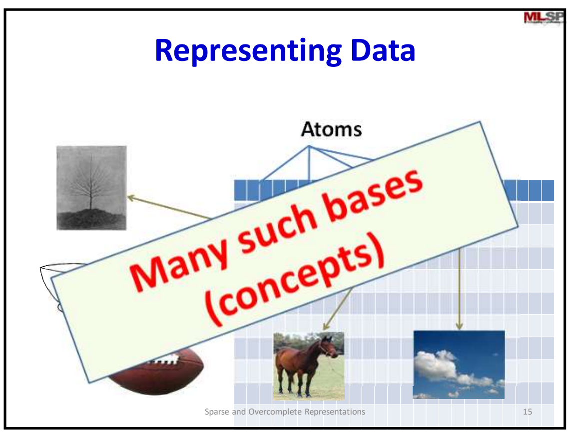
12



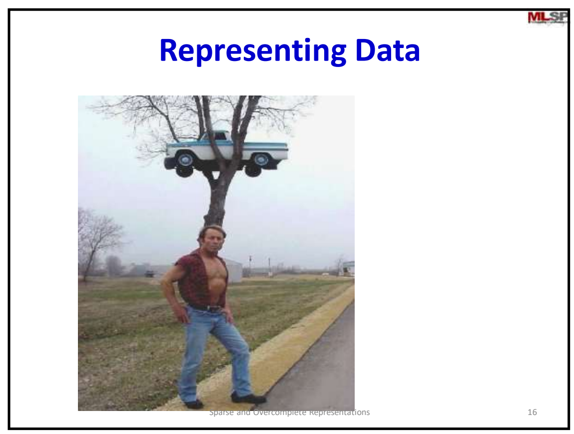
13



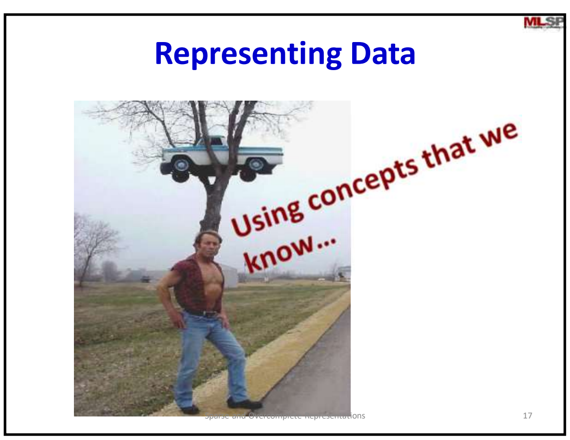
14



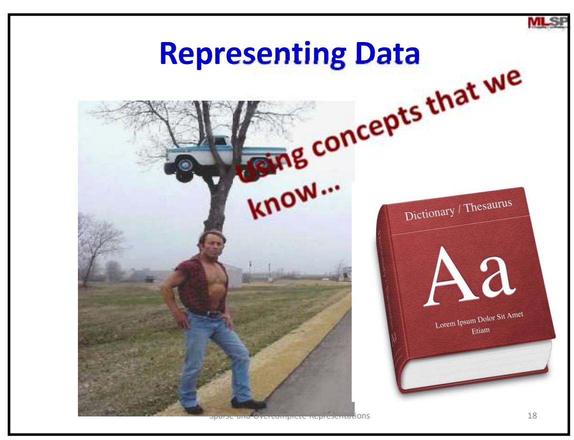
15



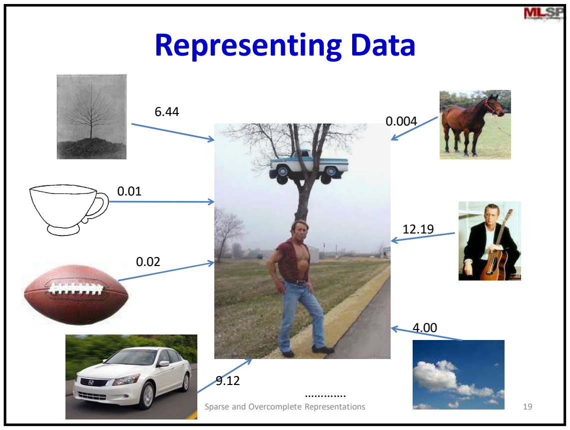
16



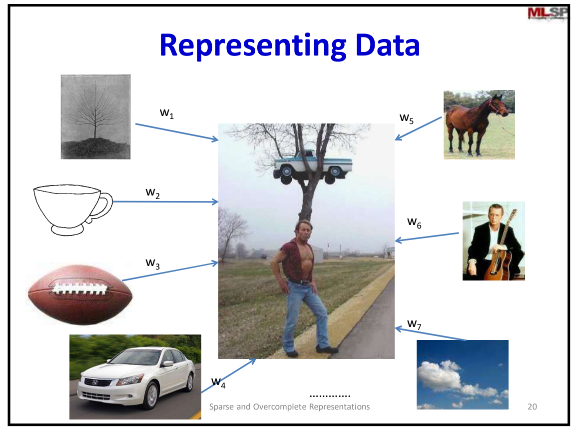
17



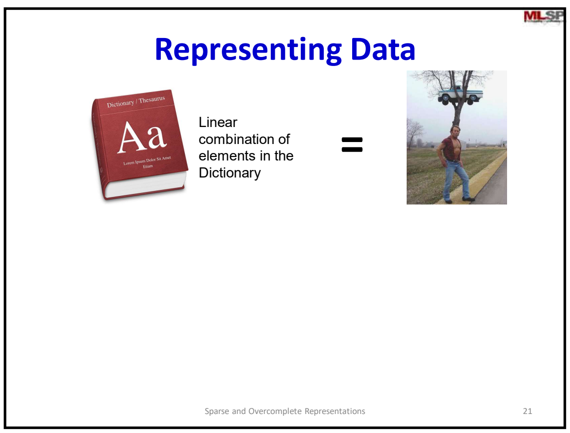
18



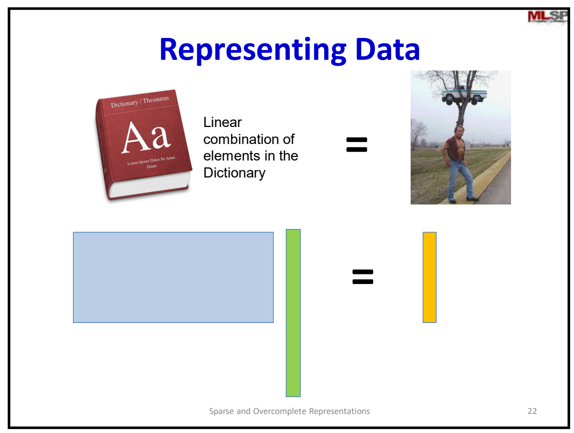
19



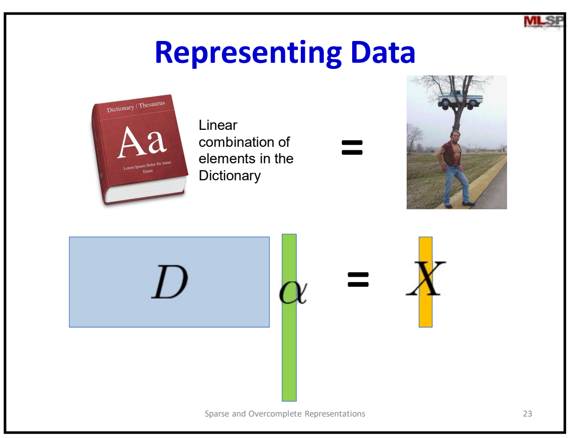
20



21



22



23

Quick Linear Algebra Refresher

- Remember, #(Basis Vectors)= #unknowns

$$D \cdot \alpha = X$$

Basis Vectors ***(from Dictionary)***

Weights

Input data

Sparse and Overcomplete Representations 24

24

Overcomplete Representations

- What is the dimensionality of the input image? (say 64x64 image)
  - **4096**
- What is the dimensionality of the dictionary? (each image = 64x64 pixels)
  - **4096 x N**

25

Overcomplete Representations

- What is the dimensionality of the input image? (say 64x64 image)
  - **4096**
- What is the dimensionality of the dictionary? (each image = 64x64 pixels)
  - **4096 x N** ???

26

Overcomplete Representations

- What is the dimensionality of the input image? (say 64x64 image)
  - **4096**
- What is the dimensionality of the dictionary? (each image = 64x64 pixels)
  - **4096 x N** VERY LARGE!!!

27

Overcomplete Representations

- What is the dimensionality of the input image? (say 64x64 image)

If N > 4096 (as it likely is)
we have an **overcomplete** representation

- What is the dimensionality of the dictionary? (each image = 64x64 pixels)
  - **4096 x N** VERY LARGE!!!

28

Overcomplete Representations

- What is the dimensionality of the input image? (say 64x64 image)

More generally:

If #(dictionary units) > dimensions of input

we have an **overcomplete** representation

- **4096 x N** VERY LARGE!!!

29

Quick Linear Algebra Refresher

- Remember, #(Basis Vectors)= #unknowns

$$D \cdot \alpha = X$$

Dictionary Units

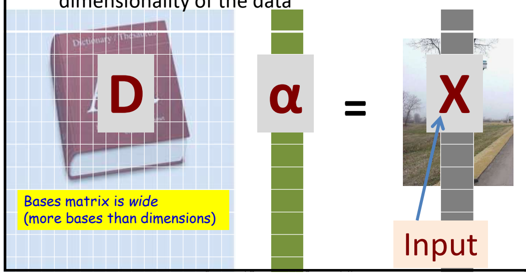
Weights

Input data

30

Dictionary based Representations

- Overcomplete "dictionary"-based representations are linear-composition-based representations with more "atomic building blocks" than the dimensionality of the data



31

Why Dictionary-based Representations?

- Dictionary based representations are semantically more meaningful
- Enable content-based description
  - Bases can capture entire structures in data
  - E.g. notes in music
  - E.g. image structures (such as faces) in images
- Enable content-based processing
  - Reconstructing, separating, denoising, manipulating speech/music signals
  - Coding, compression, etc.
- Statistical reasons: We will get to that shortly..

32

Poll 2

- Dictionary-based representations are similar to vector-quantization based representations, except that the weights vector **w** is no longer required to be one-hot
  - True
  - False
- Dictionary based representations are similar to PCA/KLT, except that the dictionary entries may exceed the dimensionality of the data in number and are not restricted to being orthogonal
  - True
  - False

33

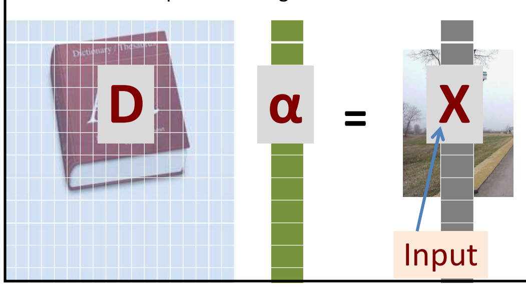
Poll 2

- Dictionary-based representations are similar to vector-quantization based representations, except that the weights vector **w** is no longer required to be one-hot
  - **True**
  - False
- Dictionary based representations are similar to PCA/KLT, except that the dictionary entries may exceed the dimensionality of the data in number and are not restricted to being orthogonal
  - **True**
  - False

34

Problems

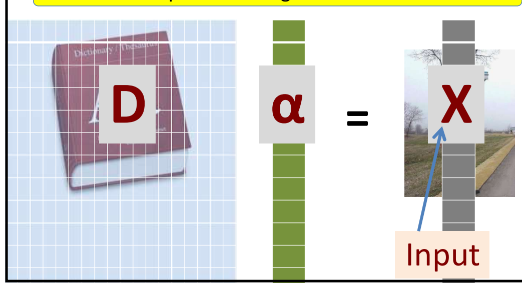
- How to obtain the dictionary
  - Which will give us meaningful representations
- How to compute the weights?



35

Problems

- How to obtain the dictionary
  - Which will give us meaningful representations
- How to compute the weights?



36

Quick Linear Algebra Refresher

- Remember, #(Basis Vectors)= #unknowns

$$D \cdot \alpha = X$$

Dictionary entries → D; Weights → α; Input data → X

When can we solve for α?

37

Quick Linear Algebra Refresher

$$D \cdot \alpha = X$$

D: full rank

D α = X — Unique solution

D α = X — We may have no exact solution

D α = X — Infinite Solutions

38

Quick Linear Algebra Refresher

$$D \cdot \alpha = X$$

D: full rank

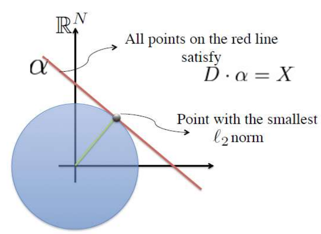
D α = X — Unique solution

D α = X — We may have no exact solution

D α = X — Our Case — Infinite Solutions

39

Using Pseudo-Inverse?



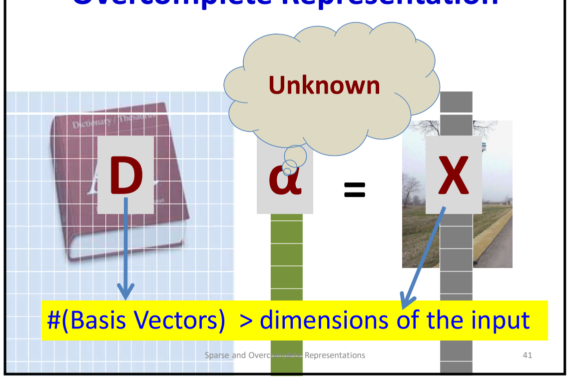
This is equivalent to

minimize $\|\alpha\|_2$ subject to $D\alpha = X$

α will generally be "dense"

40

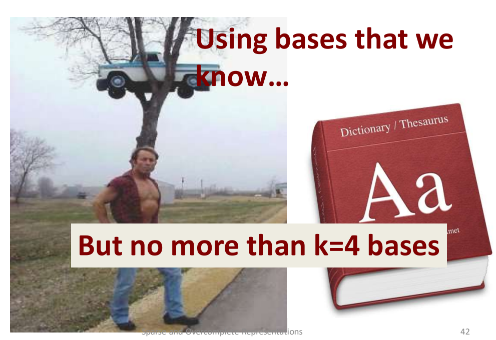
Overcomplete Representation



#(Basis Vectors) > dimensions of the input

41

Representing Data

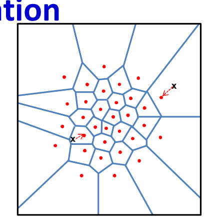


42

Alternate view: Recall quantization

$$V = \sum_i w_i d_i$$

$$V = \mathbf{D}\mathbf{w} \qquad |\mathbf{w}| = 1$$
$$|\mathbf{w}|_0 = 1$$



- d_i are the "representative" vectors of each cluster
- Restriction: only one of the w_i is 1, the rest are 0
  - $\sum_i w_i = 0$
  - $\mathbf{w}$ is unit length and one-sparse
- What if we let *more* than one entry of **w** to be non zero?

43

43

Overcompleteness and Sparsity

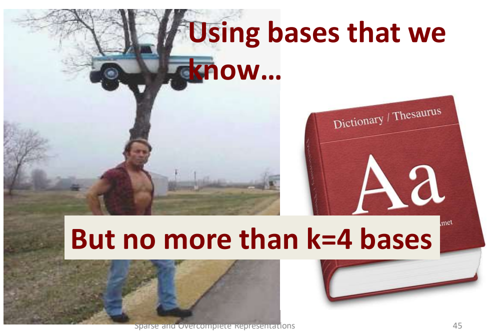
- To solve an overcomplete system of the type:

  D.α = X
- Make assumptions about the data.
- Suppose, we say that **X** is composed of no more than a fixed number (**k**) of "bases" from **D** (k ≤ dim(**X**))
  - The term "bases" is an abuse of terminology..
- Now, we can find the set of **k** bases that best fit the data point, **X**.

Sparse and Overcomplete Representations 44

44

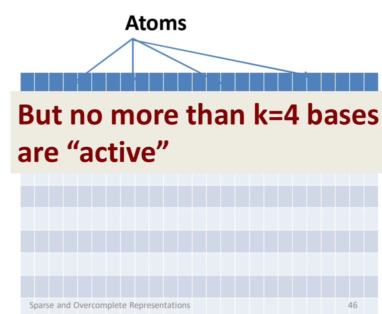
Representing Data



Sparse and Overcomplete Representations 45

45

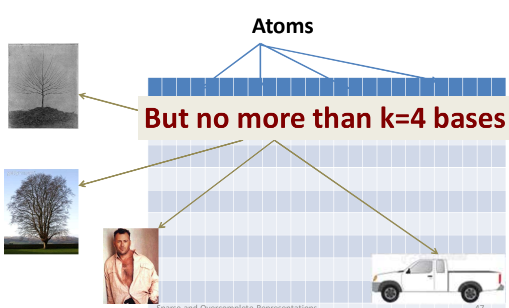
Overcompleteness and Sparsity



Sparse and Overcomplete Representations 46

46

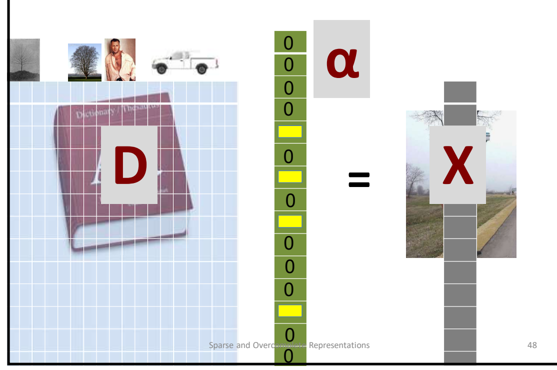
Overcompleteness and Sparsity



Sparse and Overcomplete Representations 47

47

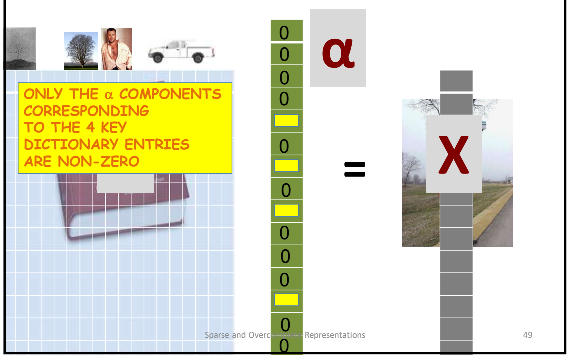
No more than 4 bases



Sparse and Overcomplete Representations 48

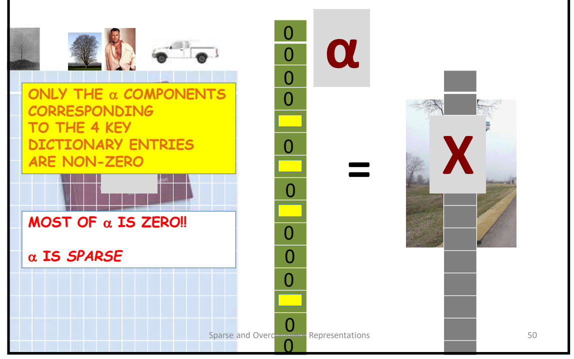
48

No more than 4 bases



49

No more than 4 bases



50

Sparsity- Definition

- *Sparse representations* are representations that account for most or all information of a signal with a linear combination of a small number of atoms.

(*from*: www.see.ed.ac.uk/~tblumens/Sparse/Sparse.html)

Sparse and Overcomplete Representations 51

51

The Sparsity Problem

- We don't really know **k**
- You are given a signal **X**
- Assuming **X** was generated using the dictionary, can we find α that generated it?

Sparse and Overcomplete Representations 52

52

The Sparsity Problem

- We want to use as few dictionary entries as possible to do this.

$$\underset{\alpha}{Min}\ \|\alpha\|_0 \quad s.t.\ \ X = \mathbf{D}\alpha$$

Sparse and Overcomplete Representations 53

53

The Sparsity Problem

- We want to use as few dictionary entries as possible to do this.

$$\underset{\alpha}{Min}\ \|\alpha\|_0 \quad s.t.\ \ X = \mathbf{D}\alpha$$

Counts the number of non-zero elements in α

Sparse and Overcomplete Representations 54

54

The Sparsity Problem

- We want to use **as few dictionary entries** as possible to do this
  - Ockham's razor: Choose the simplest explanation invoking the fewest variables

$$\underset{\underline{\alpha}}{Min}\ \|\underline{\alpha}\|_0 \quad s.t.\ \ \underline{X} = \mathbf{D}\underline{\alpha}$$

55

Poll 3

- Overcomplete representations can be indeterminate
  - True
  - False
- It is essential to impose sparsity to obtain a unique representation in terms of an overcomplete dictionary
  - True
  - False

56

Poll 3

- Overcomplete representations can be indeterminate
  - **True**
  - False
- It is essential to impose sparsity to obtain a unique representation in terms of an overcomplete dictionary
  - True
  - **False**

57

The Sparsity Problem

- We want to use as few dictionary entries as possible to do this.

$$\underset{\underline{\alpha}}{Min}\ \|\underline{\alpha}\|_0 \quad s.t.\ \ \underline{X} = \mathbf{D}\underline{\alpha}$$

How can we solve the above?

58

Obtaining Sparse Solutions

- We will look at 2 algorithms:
  - Matching Pursuit (MP)
  - Basis Pursuit (BP)

59

Matching Pursuit (MP)

- Greedy algorithm
- Finds an atom in the dictionary that best matches the input signal
- Remove the weighted value of this atom from the signal
- Again, find an atom in the dictionary that best matches the remaining signal.
- Continue till a defined stop condition is satisfied.

60

Matching Pursuit

- Find the dictionary atom that best matches the given signal.

Weight = w_1

Sparse and Overcomplete Representations 61

61

Matching Pursuit

- Remove weighted image to obtain updated signal

Find best match for this signal from the dictionary

Sparse and Overcomplete Representations 62

62

Matching Pursuit

- Find best match for updated signal

Sparse and Overcomplete Representations 63

63

Matching Pursuit

- Find best match for updated signal

Iterate till you reach a stopping condition, **norm(ResidualInputSignal) < threshold**

Sparse and Overcomplete Representations 64

64

Matching Pursuit

Algorithm Matching Pursuit
Input: Signal: $f(t)$.
Output: List of coefficients: (a_n, g_{γ_n}).
Initialization:
$Rf_1 \leftarrow f(t)$;
Repeat
find $g_{\gamma_n} \in D$ with maximum inner product $< Rf_n, g_{\gamma_n} >$;
$a_n \leftarrow < Rf_n, g_{\gamma_n} >$;
$Rf_{n+1} \leftarrow Rf_n - a_n g_{\gamma_n}$;
$n \leftarrow n+1$;
Until stop condition (for example: $\|Rf_n\| < threshold$)

From http://en.wikipedia.org/wiki/Matching_pursuit

Sparse and Overcomplete Representations 65

65

Matching Pursuit

- Problems ???

Sparse and Overcomplete Representations 66

66

Matching Pursuit

- Main Problem
  - Computational complexity
  - The entire dictionary has to be searched at every iteration

Comparing MP and BP

Matching Pursuit	Basis Pursuit
Hard thresholding (remember the equations)	
Greedy optimization at each step	
Weights obtained using greedy rules	

Basis Pursuit (BP)

- Remember,

$$\underset{\underline{\alpha}}{Min} \; \|\underline{\alpha}\|_0 \quad s.t. \;\; \underline{X} = \mathbf{D}\underline{\alpha}$$

Basis Pursuit

- Remember,

$$\underset{\underline{\alpha}}{Min} \; \|\underline{\alpha}\|_0 \quad s.t. \;\; \underline{X} = \mathbf{D}\underline{\alpha}$$

In the general case, this is intractable

Basis Pursuit

- Remember,

$$\underset{\underline{\alpha}}{Min} \; \|\underline{\alpha}\|_0 \quad s.t. \;\; \underline{X} = \mathbf{D}\underline{\alpha}$$

In the general case, this is intractable

Requires combinatorial optimization

Basis Pursuit

- Replace the intractable expression by an expression that is solvable

$$\underset{\underline{\alpha}}{Min} \; \|\underline{\alpha}\|_1 \quad s.t. \;\; \underline{X} = \mathbf{D}\underline{\alpha}$$

Basis Pursuit

- Replace the intractable expression by an expression that is solvable

$$\underset{\alpha}{Min}\ \|\alpha\|_1 \quad s.t.\ \ X = \mathbf{D}\alpha$$

This will provide identical solutions when $\mathbf{D}$ obeys the ***Restricted Isometry Property***.

73

Basis Pursuit

- Replace the intractable expression by an expression that is solvable

$$\underset{\alpha}{Min}\ \|\alpha\|_1 \quad s.t.\ \ X = \mathbf{D}\alpha$$

Objective

Constraint

74

Basis Pursuit

- We can formulate the optimization term as:

$$\underset{\alpha}{Min}\ \{\|X - \mathbf{D}\alpha\|^2 + \lambda\|\alpha\|_1\}$$

Constraint Objective

75

Basis Pursuit

- We can formulate the optimization term as:

$$\underset{\alpha}{Min}\ \{\|X - \mathbf{D}\alpha\|^2 + \lambda\|\alpha\|_1\}$$

λ is a penalty term on the non-zero elements and promotes sparsity

76

Basis Pursuit

Equivalent to *LASSO*; for more details, see this paper by Tibshirani
http://www-stat.stanford.edu/~tibs/ftp/lasso.ps

$$\underset{\alpha}{Min}\ \{\|X - \mathbf{D}\alpha\|^2 + \lambda\|\alpha\|_1\}$$

λ is a penalty term on the non-zero elements and promotes sparsity

77

Basis Pursuit

$$\underset{\alpha}{Min}\{\|X - \mathbf{D}\alpha\|^2 + \lambda\|\alpha\|_1\}$$

$$\frac{\partial\|\alpha\|_1}{\partial\alpha_j} = \begin{cases} +1 & \text{at } \alpha_j > 0 \\ [-1,1] & \text{at } \alpha_j = 0 \\ -1 & \text{at } \alpha_j < 0 \end{cases}$$

- $\|\alpha\|_1$ is not differentiable at $\alpha_j = 0$
- Gradient of $\|\alpha\|_1$ for gradient descent update
- At optimum, following conditions hold

$$\nabla_j\|X - \mathbf{D}\alpha\|^2 + \lambda sign(\alpha_j) = 0, \quad \text{if } |\alpha_j| > 0$$

$$\nabla_j\|X - \mathbf{D}\alpha\|^2 \le \lambda, \quad \text{if } \alpha_j = 0$$

78

Basis Pursuit

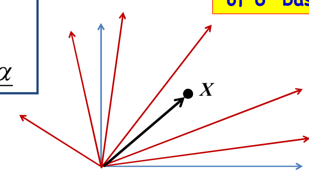
- There are efficient ways to solve the LASSO formulation.
  - http://web.stanford.edu/~hastie/glmnet_matlab/
- Simplest solution: Coordinate descent algorithms
  - On webpage..

79

L_1 vs L_0

$$\min_{\alpha} \|\alpha\|_0 \quad s.t.\ \underline{X} = \mathbf{D}\underline{\alpha}$$

Overcomplete set of 6 "bases"



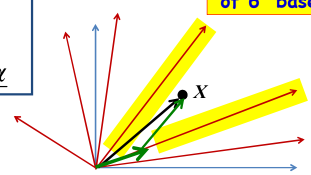
- **L_0 minimization**
  - Two-sparse solution
  - ANY pair of bases can explain X with 0 error

80

L_1 vs L_0

$$\min_{\alpha} \|\alpha\|_1 \quad s.t.\ \underline{X} = \mathbf{D}\underline{\alpha}$$

Overcomplete set of 6 "bases"



- **L_1 minimization**
  - Two-sparse solution
  - All else being equal, the two closest bases are chosen

81

Comparing MP and BP

Matching Pursuit	Basis Pursuit
Hard thresholding	Soft thresholding
(remember the equations)	
Greedy optimization at each step	Global optimization
Weights obtained using greedy rules	Can force N-sparsity with appropriately chosen weights

82

General Formalisms

- L_0 minimization $\quad \min_{\alpha} \|\alpha\|_0 \quad s.t.\ \underline{X} = \mathbf{D}\underline{\alpha}$
- L_0 constrained optimization $\quad \min_{\alpha} \|\underline{X} - \mathbf{D}\underline{\alpha}\|_2^2 \quad s.t.\ \|\underline{\alpha}\|_0 < C$
- L_1 minimization $\quad \min_{\alpha} \|\alpha\|_1 \quad s.t.\ \underline{X} = \mathbf{D}\underline{\alpha}$
- L_1 constrained optimization $\quad \min_{\alpha} \|\underline{X} - \mathbf{D}\underline{\alpha}\|_2^2 \quad s.t.\ \|\underline{\alpha}\|_1 < C$

83

Many Other Methods..

- Iterative Hard Thresholding (IHT)
- CoSAMP
- OMP
- …

84

Poll 4

- Which of the following are valid ways of obtaining a sparse representation w
  - Minimize |w|_0 while constraining X = Dw
  - Minimize |w|_0 while constraining X <= Dw
  - Minimize ||X-Dw||^2 + lambda*|w|_1
  - Minimize |w|_2 while constraining X <= Dw

85

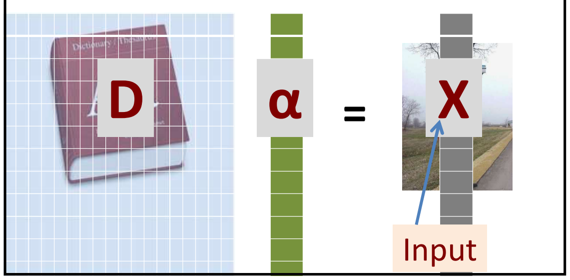
Poll 4

- Which of the following are valid ways of obtaining a sparse representation w
  - **Minimize |w|_0 while constraining X = Dw**
  - Minimize |w|_0 while constraining X <= Dw
  - **Minimize ||X-Dw||^2 + lambda*|w|_1**
  - Minimize |w|_2 while constraining X <= Dw

86

Problems

- How to obtain the dictionary
  - Which will give us meaningful representations
- How to compute the weights?



87

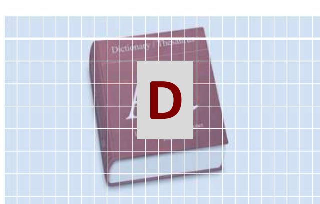
Trivial Solution

- D = Training data
- Impractical in most situations
  - Popular approach: sample random vectors from training data

88

Dictionaries: Compressive Sensing

- Just random vectors!



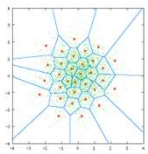
89

More Structured ways of Constructing Dictionaries

- Dictionary entries must be structurally “meaningful”
  - Represent true compositional units of data
- Have already encountered two ways of building dictionaries
  - NMF for non-negative data
  - K-means ..

90

K-Means for Composing Dictionaries



Train the codebook from training data using K-means

- Every vector is approximated by the centroid of the cluster it falls into
- Cluster means are "codebook" entries
  - Dictionary entries
  - Also compositional units the compose the data

91

K-Means for Dictionaries

Each column is a codeword (centroid) from the codebook

D **α** 0 0 0 0 1 0 0 0 0 0 0 0 0 0

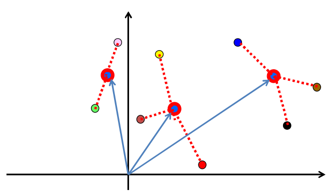
X

- α *must* be 1 sparse
- Only α entry must 1

$||\alpha||_0 = 1$ $||\alpha||_1 = 1$

92

K-Means



- Learn Codewords to minimize the total squared length of the training vectors from the closest codeword

93

Length-unconstrained K-Means for Dictionaries

Each column is a codeword (centroid) from the codebook

D **α** 0 0 0 0 3 0 0 0 0 0 0 0 0 0

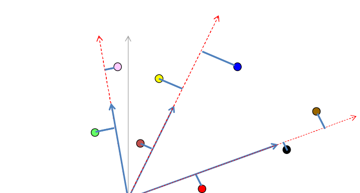
X

- α *must* be 1 sparse
- No restriction on α value

$||\alpha||_0 = 1$

94

SVD K-Means

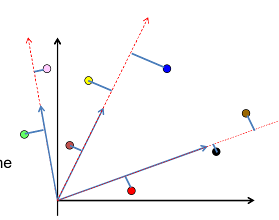


- Learn Codewords to minimize the total squared *projection error* of the training vectors from the closest codeword

95

SVD K–means

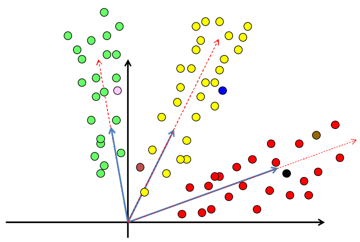
1. Initialize a set of (unit-length) centroids randomly
2. For each data point x, find the projection from the centroid for each cluster
   - $p_{cluster} = |x^T m_{cluster}|$
3. Put data point in the cluster of the closest centroid
   - Cluster for which $p_{cluster}$ is *maximum*
4. When all data points are clustered, recompute centroids



$$m_{cluster} = \text{Principal Eigenvector}(\{x \mid x \in cluster\})$$

96

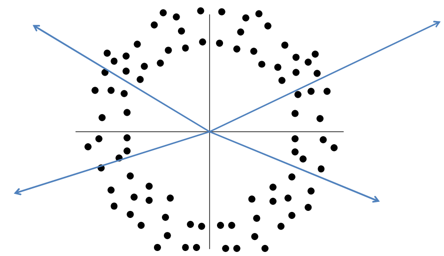
Problem



- Only represents *Radial* patterns

97

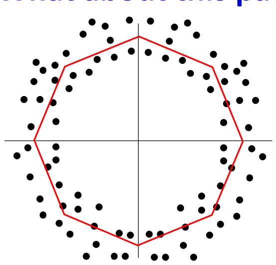
What about this pattern?



- Dictionary entries that represent radial patterns will not capture this structure
  - 1-sparse representations will not do

98

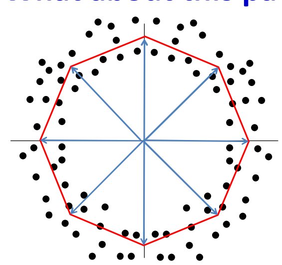
What about this pattern?



- We need ***AFFINE*** *patterns*

99

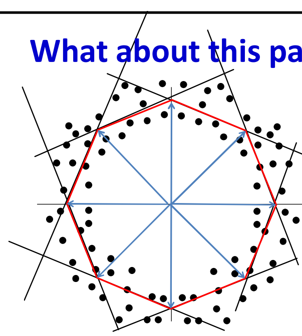
What about this pattern?



- We need ***AFFINE*** *patterns*
- *Each vector is modeled by a linear combination of K (here 2) bases*

100

What about this pattern?



Every line is a (constrained) combination of two bases

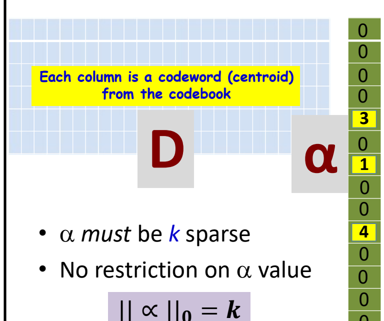
2-sparse

Constraint:
Line = a.b_1 + (1-a)b_2

- We need ***AFFINE*** *patterns*
- *Each vector is modeled by a linear combination of K (here 2) bases*

101

Codebooks for K sparsity?



- α *must* be *k* sparse
- No restriction on α value

$|| \alpha ||_0 = k$

102

Formalizing

Given training data

$$\{X_1, X_2, ..., X_T\}$$

We want to find a dictionary D, such that

$$D\alpha_i = X_i$$

With α_i sparse

103

Formalizing

Two objectives:
- Approximation $\|D\alpha_i - X_i\|$
- Sparsity in coefficients $\|\alpha_i\|_1$

$$\min_{D,\alpha_i} \sum_{i=1}^{T} \|X_i - D\alpha_i\|^2 + \|\alpha_i\|_1$$

NON-Convex!!!

104

An iterative method

- Given D, estimate α_i to get sparse solution
  - We can use any method

$$\min_{\alpha_i} \sum_{i=1}^{T} \|X_i - D\alpha_i\|^2 + \|\alpha_i\|_1$$

- Given α_i, estimate D

$$\min_{D} \sum_{i=1}^{T} \|X_i - D\alpha_i\|^2$$

Difficult!

105

K SVD

- Initialize Codebook

D =

1. For every vector, compute K-sparse alphas
   – Using any pursuit algorithm

α =

0	0	2	0	1
1	0	0	0	0
0	0	0	1	0
0	7	0	0	0
0	0	0	1	0
3	0	0	0	0
0	5	0	0	1
0	0	1	0	2

106

K-SVD

2. For each codeword (k):
- For each vector x that used k
  - Subtract the contribution of all other codewords to obtain $e_k(x)$
    - Codeword-specific residual
- Compute the principal Eigen vector of $\{e_k(x)\}$
3. Return to step 1

D = (D_j j != 1)

α = ($\alpha_j(x)$ j != 1)

0	0	2	0	1
1	0	0	0	0
0	0	0	1	0
0	7	0	0	0
0	0	0	1	0
3	0	0	0	0
0	5	0	0	1
0	0	1	0	2

$$e_k(x) = x - \sum_{j \neq k} \alpha_j D_j$$

107

K-SVD

D

- Termination of each iteration: Updated dictionary
- Conclusion: A dictionary where any data vector can be composed of at most K dictionary entries
  – More generally, sparse composition

108

K-SVD algorithm (skip)

Initialization : Set the random normalized dictionary matrix $\mathbf{D}^{(0)} \in \mathbb{R}^{n\times K}$. Set $J = 1$.
Repeat until convergence,
Sparse Coding Stage: Use any pursuit algorithm to compute $\mathbf{x}_i$ for $i = 1, 2, \ldots, N$

$$\min_{\mathbf{x}} \{\|\mathbf{y}_i - \mathbf{D}\mathbf{x}\|_2^2\} \quad \text{subject to} \quad \|\mathbf{x}\|_0 \le T_0.$$

Codebook Update Stage: For $k = 1, 2, \ldots, K$

- Define the group of examples that use $\mathbf{d}_k$, $\omega_k = \{i \mid 1 \le i \le N,\ \mathbf{x}_i(k) \ne 0\}$.
- Compute
$$\mathbf{E}_k = \mathbf{Y} - \sum_{j\ne k} \mathbf{d}_j \mathbf{x}^j,$$
- Restrict $\mathbf{E}_k$ by choosing only the columns corresponding to those elements that initially used $\mathbf{d}_k$ in their representation, and obtain $\mathbf{E}_k^R$.
- Apply SVD decomposition $\mathbf{E}_k^R = \mathbf{U}\Delta\mathbf{V}^T$. Update: $\mathbf{d}_k = \mathbf{u}_1$, $\mathbf{x}_R^k = \Delta(1,1)\cdot\mathbf{v}_1$

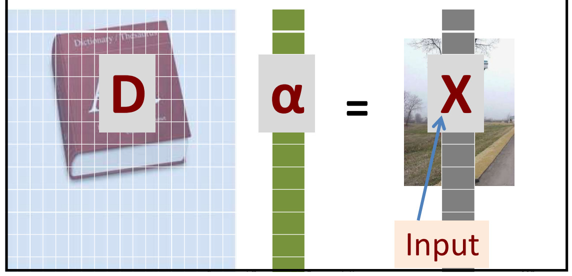
Set $J = J + 1$.

Sparse and Overcomplete Representations 109

109

Problems

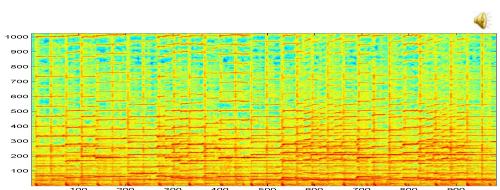
- How to obtain the dictionary
  - Which will give us meaningful representations
- How to compute the weights?



Sparse and Overcomplete Representations 110

110

So how does that work

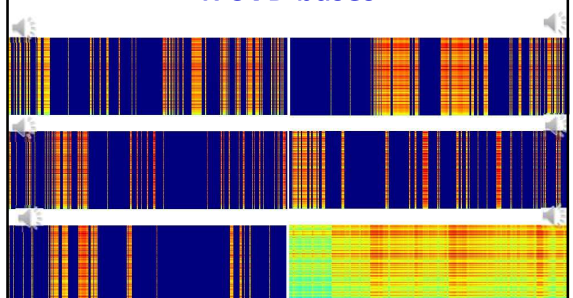


- In case you forgot this music…
- 975 vectors (1025 dimensions)
- N=12, K=5

Sparse and Overcomplete Representations 111

111

K-SVD bases



Sparse and Overcomplete Representations 112

112

Applications of Sparse Representations

- Many many applications
  - Signal representation
  - Statistical modelling
  - ..
  - We've seen one: Compressive sensing
- Another popular use
  - **Denoising**

Sparse and Overcomplete Representations 113

113

Denoising

- As the name suggests, remove noise!

Sparse and Overcomplete Representations 114

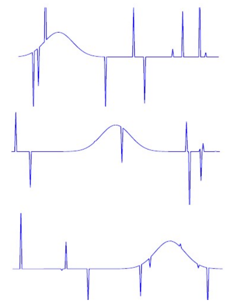
114

Denoising

- As the name suggests, remove noise!
- We will look at image denoising as an example

115

A toy example



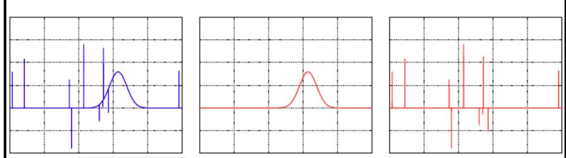
116

A toy example

$$D = [I \ G]$$

I Identity matrix

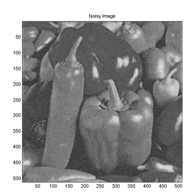
G Translation of a Gaussian pulse



117

Image Denoising

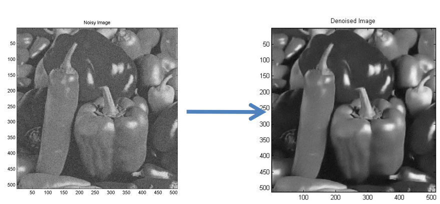
- Here's what we want



118

Image Denoising

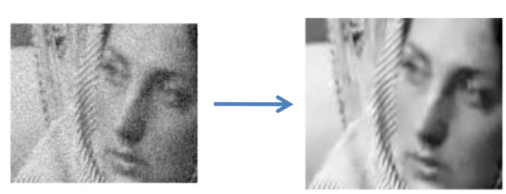
- Here's what we want



119

Image Denoising

- Here's what we want



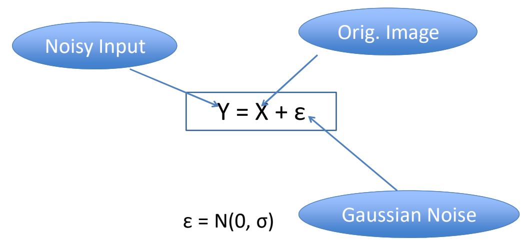
120

The Image Denoising Problem

- Given an image
- Remove Gaussian additive noise from it

121

Image Denoising



122

Image Denoising

- Remove the noise from **Y**, to obtain **X** as best as possible.

123

Image Denoising

- Remove the noise from **Y**, to obtain **X** as best as possible
- Using sparse representations over learned dictionaries

124

Image Denoising

- Remove the noise from **Y**, to obtain **X** as best as possible
- Using sparse representations over learned dictionaries
- We will *learn* the dictionaries

125

Image Denoising

- Remove the noise from **Y**, to obtain **X** as best as possible
- Using sparse representations over learned dictionaries
- We will *learn* the dictionaries
- What data will we use? *The corrupted image itself!*

126

Image Denoising

- We use the data to be denoised to learn the dictionary.
- Training and denoising become an iterated process.
- We use image patches of size √n x √n pixels (i.e. if the image is 64x64, patches are 8x8)

Sparse and Overcomplete Representations 127

127

Image Denoising

- The data dictionary D
  - Size = n x k (k > n)
  - This is known and fixed, to start with
  - Every image patch can be sparsely represented using D

Sparse and Overcomplete Representations 128

128

Image Denoising

- Recall our equations from before.
- We want to find α so as to minimize the value of the equation below:

$$\underset{\underline{\alpha}}{Min}\ \{\|\underline{X}-\mathbf{D}\underline{\alpha}\|^2+\lambda\|\underline{\alpha}\|_0\}$$

$$\underset{\underline{\alpha}}{Min}\{\|\underline{X}-\mathbf{D}\underline{\alpha}\|^2+\lambda\|\underline{\alpha}\|_1\}$$

Sparse and Overcomplete Representations 129

129

Image Denoising

$$\underset{\underline{\alpha}}{Min}\ \{\|\underline{X}-\mathbf{D}\underline{\alpha}\|^2+\lambda\|\underline{\alpha}\|_1\}$$

- In the above, *X* is a patch.

Sparse and Overcomplete Representations 130

130

Image Denoising

$$\underset{\underline{\alpha}}{Min}\ \{\|\underline{X}-\mathbf{D}\underline{\alpha}\|^2+\lambda\|\underline{\alpha}\|_1\}$$

- In the above, *X* is a patch.
- If the larger image is fully expressed by the every patch in it, how can we go from patches to the image?

Sparse and Overcomplete Representations 131

131

Image Denoising

$$\underset{\underline{\alpha_{ij}},X}{Min}\ \{\mu\|\underline{X}-Y\|_2^2+\sum_{ij}\left\|\underline{R_{ij}X}-\mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 +\sum_{ij}\lambda_{ij}\left\|\underline{\alpha_{ij}}\right\|_0\}$$

Sparse and Overcomplete Representations 132

132

Image Denoising

$$\underset{\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

(X – Y) is the error between the input and denoised image. μ is a penalty on the error.

133

Image Denoising

$$\underset{\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

Error bounding in each patch
- R_{ij} selects the (ij)th patch
- Terms in summation = no. of patches

134

Image Denoising

$$\underset{\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

λ forces sparsity

135

Image Denoising

- But, we don't "***know***" our dictionary D.
- We want to estimate D as well.

136

Image Denoising

- But, we don't "***know***" our dictionary D.
- We want to estimate D as well.

$$\underset{D,\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

We can use the previous equation itself!!!

137

Image Denoising

$$\underset{D,\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

How do we estimate all 3 at once?

138

Image Denoising

$$\underset{D,\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2$$
$$+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

How do we estimate all 3 at once?

We cannot estimate them at the same time!

139

Image Denoising

$$\underset{D,\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2$$
$$+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

How do we estimate all 3 at once?

Fix 2, and find the optimal 3rd.

140

Image Denoising

$$\underset{D,\alpha_{ij},X}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2$$
$$+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

Initialize X = Y

141

Image Denoising

$$\underset{\alpha_{ij}}{Min}\ \{\mu\|X-Y\|_2^2+\sum_{ij}\|R_{ij}X-\mathbf{D}\alpha_{ij}\|_2^2$$
$$+\sum_{ij}\lambda_{ij}\|\alpha_{ij}\|_0\}$$

0

Initialize X = Y, initialize D

You know how to solve the remaining portion for α – MP, BP!

142

Image Denoising

- Now, update the dictionary D.
- Update D one column at a time, following the K-SVD algorithm
- K-SVD maintains the sparsity structure

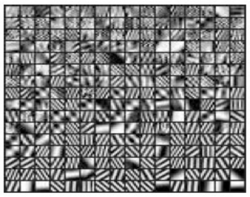
143

Image Denoising

- Now, update the dictionary D.
- Update D one column at a time, following the K-SVD algorithm
- K-SVD maintains the sparsity structure
- Iteratively update α and D

144

Image Denoising



Learned Dictionary for Face Image denoising

From: M. Elad and M. Aharon, *Image denoising via learned dictionaries and sparse representation*, CVPR, 2006.

145

Image Denoising

$$\underset{\underline{X}}{Min}\ \{\mu\|\underline{X}-Y\|_2^2+\sum_{ij}\|\underline{R_{ij}X}-\mathbf{D}\underline{\alpha}_{ij}\|_2^2 + \sum_{ij}\lambda_{ij}\|\underline{\alpha_{ij}}\|_0\}$$

→ Const. wrt X

We know D and α

The quadratic term above has a closed-form solution

146

Image Denoising

$$\underset{\underline{X}}{Min}\ \{\mu\|\underline{X}-Y\|_2^2+\sum_{ij}\|\underline{R_{ij}X}-\mathbf{D}\underline{\alpha}_{ij}\|_2^2 + \sum_{ij}\lambda_{ij}\|\underline{\alpha_{ij}}\|_0\}$$

→ Const. wrt X

We know D and α

$$X=(\mu I+\sum_{ij}R_{ij}^TR_{ij})^{-1}(\mu Y+\sum_{ij}R_{ij}^TD\alpha_{ij})$$

147

Image Denoising

- Summarizing… We wanted to obtain 3 things

148

Image Denoising

- Summarizing… We wanted to obtain 3 things

- Weights **α**
- Dictionary **D**
- Denoised Image **X**

149

Image Denoising

- Summarizing… We wanted to obtain 3 things

- Weights **α –** Your favorite pursuit algorithm
- Dictionary **D –** Using K-SVD
- Denoised Image **X**

150

Image Denoising

- Summarizing… We wanted to obtain 3 things

➢ Weights **α** – Your favorite pursuit algorithm
➢ Dictionary **D** – Using K-SVD
➢ Denoised Image **X**

Iterating

151

Image Denoising

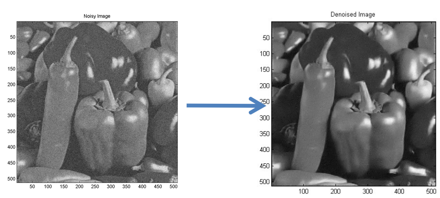
- Summarizing… We wanted to obtain 3 things

➢ Weights **α**
➢ Dictionary **D**
➢ Denoised Image **X**- Closed form solution

152

Image Denoising

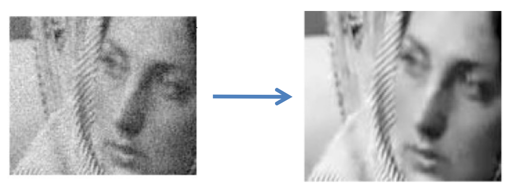
- Here's what we want



153

Image Denoising

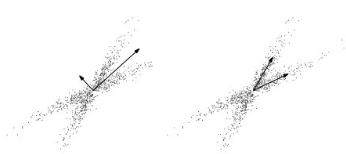
- Here's what we want



154

Comparing to Other Techniques

Non-Gaussian data



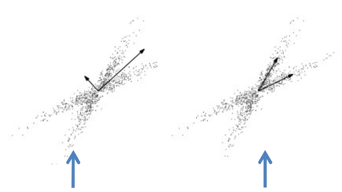
PCA of ICA Which is which?

Images from Lewicki and Sejnowski, *Learning Overcomplete Representations*, 2000.

155

Comparing to Other Techniques

Non-Gaussian data



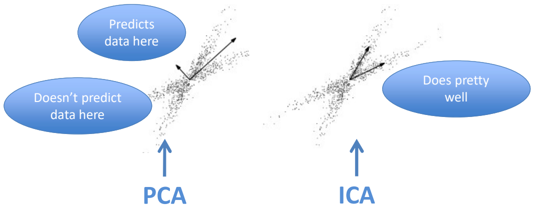
PCA ICA

Images from Lewicki and Sejnowski, *Learning Overcomplete Representations*, 2000.

156

Comparing to Other Techniques

Non-Gaussian data

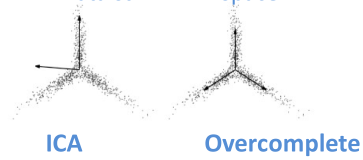


Images from Lewicki and Sejnowski, *Learning Overcomplete Representations*, 2000.

157

Comparing to Other Techniques

Data still in 2-D space



Doesn't capture the underlying representation, which Overcomplete representations can do…

158

Summary

- Overcomplete representations can be more powerful than component analysis techniques.
- Dictionary can be learned from data.
- Relative advantages and disadvantages of the pursuit algorithms.

159